Benefits of Bias: Towards Better Characterization of Network Sampling

Arun S. Maiya
Dept. of Computer Science
University of Illinois at Chicago
arun@maiya.net

Tanya Y. Berger-Wolf
Dept. of Computer Science
University of Illinois at Chicago
tanyabw@cs.uic.edu

ABSTRACT

From social networks to P2P systems, network sampling arises in many settings. We present a detailed study on the nature of biases in network sampling strategies to shed light on how best to sample from networks. We investigate connections between specific biases and various measures of structural representativeness. We show that certain biases are, in fact, beneficial for many applications, as they "push" the sampling process towards inclusion of desired properties. Finally, we describe how these sampling biases can be exploited in several, real-world applications including disease outbreak detection and market research.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms

Algorithms; Experimentation, Measurement

Keywords

sampling, bias, social network analysis, complex networks, graph mining, link mining, online sampling, crawling

1. INTRODUCTION AND MOTIVATION

We present a detailed study on the nature of biases in network sampling strategies to shed light on how best to sample from networks. A *network* is a system of interconnected entities typically represented mathematically as a graph: a set of vertices and a set of edges among the vertices. Networks are ubiquitous and arise across numerous and diverse domains. For instance, many Web-based social media, such as online social networks, produce large amounts of data on interactions and associations among individuals. Mobile phones and location-aware devices produce copious amounts of data on both communication patterns and physical proximity between people. In the domain of biology also, from neurons to proteins to food webs, there is now access to large networks of associations among various entities and a need to analyze and understand these data.

With advances in technology, pervasive use of the Internet, and the proliferation of mobile phones and location-aware devices, networks under study today are not only substantially larger than those in the past, but sometimes exist in a decentralized form (e.g. the network of blogs or the Web itself). For many networks, their global structure is not fully visible to the public and can only be accessed through "crawls" (e.g. online social networks). These factors can make it prohibitive to analyze or even access these networks in their entirety. How, then, should one proceed in analyzing and mining these network data? One approach to addressing these issues is *sampling*: inference using small subsets of nodes and links from a network.

From epidemiological applications [13] to Web crawling [7] and P2P search [47], network sampling arises across many different settings. In the present work, we focus on a particular line of investigation that is concerned with constructing samples that match critical structural properties of the original network. Such samples have numerous applications in data mining and information retrieval. In [29], for example, structurally-representative samples were shown to be effective in inferring network protocol performance in the larger network and significantly improving the efficiency of protocol simulations. In Section 7, we discuss several additional applications. Although there have been a number of recent strides in work on network sampling (e.g. [3,25,29,34]), there is still very much that requires better and deeper understanding. Moreover, many networks under analysis, although treated as complete, are, in fact, *samples* due to limitations in data collection processes. Thus, a more refined understanding of network sampling is of general importance to network science. Towards this end, we conduct a detailed study on *network sampling biases*. There has been a recent spate of work focusing on *problems* that arise from network sampling biases including how and why biases should be avoided [1, 15, 20, 30, 46, 47]. Our work differs from much of this existing literature in that, for the first time in a comprehensive manner, we examine network sampling bias as an *asset to be exploited*. We argue that biases of certain sampling strategies can be advantageous if they "push" the sampling process towards inclusion of specific properties of interest.[1] Our main aim in the present work is to identify and understand the connections between specific sampling biases and

KDD'11, August 21–24, 2011, San Diego, California, USA.

[1]This is similar to the role of bias in stratified sampling in classical statistics.

specific definitions of structural representativeness, so that these biases can be leveraged in practical applications.

Summary of Findings. We conduct a detailed investigation of network sampling biases. We find that bias towards high *expansion* (a concept from expander graphs) offers several unique advantages over other biases such as those toward high degree nodes. We show both empirically and analytically that such an expansion bias "pushes" the sampling process towards new, undiscovered clusters and the discovery of wider portions of the network. In other analyses, we show that a simple sampling process that selects nodes with many connections from those already sampled is often a reasonably good approximation to directly sampling high degree nodes and locates well-connected (i.e. high degree) nodes significantly faster than most other methods. We also find that the breadth-first search, a widely-used sampling and search strategy, is surprisingly among the most dismal performers in terms of both discovering the network and accumulating critical, well-connected nodes. Finally, we describe ways in which some of our findings can be exploited in several important applications including disease outbreak detection and market research. A number of these aforementioned findings are surprising in that they are in stark contrast to conventional wisdom followed in much of the existing literature (e.g. [2, 13, 30, 41, 42]).

2. RELATED WORK

Not surprisingly, network sampling arises across many diverse areas. Here, we briefly describe some of these different lines of research.

Network Sampling in Classical Statistics. The concept of sampling networks first arose to address scenarios where one needed to study hidden or difficult-to-access populations (e.g. illegal drug users, prostitutes). For recent surveys, one might refer to [19, 28]. The work in this area focuses almost exclusively on acquiring unbiased estimates related to variables of interest attached to each network node. The present work, however, focuses on inferring properties related to the *network itself* (many of which are not amenable to being fully captured by simple attribute frequencies). Our work, then, is much more closely related to *representative subgraph sampling*.

Representative Subgraph Sampling. In recent years, a number of works have focused on *representative subgraph sampling*: constructing samples in such a way that they are condensed representations of the original network (e.g. [3, 25, 29, 32, 34]). Much of this work focuses on how best to produce a "universal" sample representative of *all* structural properties in the original network. By contrast, we subscribe to the view that no single sampling strategy may be appropriate for all applications. Thus, our aim, then, is to better understand the *biases* in specific sampling strategies to shed light on how best to leverage them in practical applications.

Unbiased Sampling. There has been a relatively recent spate of work (e.g. [20, 22, 47]) that focuses on constructing uniform random samples in scenarios where nodes cannot be easily drawn randomly (e.g. settings such as the Web where nodes can only be accessed through crawls). These strategies, often based on modified random walks, have been shown to be effective for various frequency estimation problems (e.g. inferring the proportion of pages of a certain language in a Web graph [22]). However, as mentioned above, the present work focuses on using samples to infer structural (and functional) properties of the *network itself*. In this regard, we found these unbiased methods to be less effective during preliminary testing. Thus, we do not consider them and instead focus our attention on other more appropriate sampling strategies (such as those mentioned in *representative subgraph sampling*).

Studies on Sampling Bias. Several studies have investigated *biases* that arise from various sampling strategies (e.g. [1, 15, 30, 31, 46]). For instance, [46] showed that, under the simple sampling strategy of picking nodes at random from a scale-free network (i.e. a network whose degree distribution follows the power law), the resultant subgraph sample will *not* be scale-free. The authors of [1, 31] showed the converse is true under traceroute sampling. Virtually all existing results on network sampling bias focus on its negative aspects. By contrast, we focus on the *advantages* of certain biases and ways in which they can be exploited in network analysis.

Property Testing. Work on sampling exists in the fields of combinatorics and graph theory and is centered on the notion of *property testing* in graphs [38]. Properties such as those typically studied in graph theory, however, may be less useful for the analysis of *real-world* networks (e.g. the exact meaning of, say, k-colorability [38] within the context of a social network is unclear). Nevertheless, theoretical work on property testing in graphs is excellently surveyed in [38].

Other Areas. Decentralized search (e.g. searching unstructured P2P networks) and Web crawling can both be framed as network sampling problems, as both involve making decisions from subsets of nodes and links from a larger network. Indeed, network sampling itself can be viewed as a problem of information retrieval, as the aim is to seek out a subset of nodes that either individually or collectively match some criteria of interest. Several of the sampling strategies we study in the present work, in fact, are graph search algorithms (e.g. breadth-first search). Thus, a number of our findings discussed later have implications for these research areas (e.g. see [39]). For reviews on decentralized search both in the contexts of complex networks and P2P systems, one may refer to [27] and [49], respectively. For examples of connections between Web crawling and network sampling, see [7, 9, 42].

3. PRELIMINARIES

3.1 Notations and Definitions

We now briefly describe some notations and definitions used throughout this paper.

Definition 1. $G = (V, E)$ is a *network* or *graph* where V is set of vertices and $E \subseteq V \times V$ is a set of edges.

Definition 2. A *sample* S is a subset of vertices, $S \subset V$.

Definition 3. $N(S)$ is the *neighborhood* of S if $N(S) = \{w \in V - S : \exists v \in S \; s.t. \; (v, w) \in E\}$.

Definition 4. G_S is the *induced subgraph* of G based on the sample S if $G_S = (S, E_S)$ where the vertex set is $S \subset V$ and the edge set is $E_S = (S \times S) \cap E$. The induced subgraph of a sample may also be referred to as a *subgraph sample*.

3.2 Datasets

We study sampling biases in a total of twelve different networks: a power grid (PowerGrid [50]), a Wikipedia voting network (WikiVote [33]), a PGP trust network (PGP [6]), a citation network (HEPTh [33]), an email network (Enron [33]), two co-authorship networks (CondMat [33] and AstroPh [33]), two P2P file-sharing networks (Gnutella04 [33] and Gnutella31 [33]), two online social networks (Epinions [33] and Slashdot [33]), and a product co-purchasing network (Amazon [33]). These datasets were chosen to represent a rich set of diverse networks from different domains. This diversity allows a more comprehensive study of network sampling and thorough assessment of the performance of various sampling strategies in the face of varying network topologies. Table 1 shows characteristics of each dataset. All networks are treated as undirected and unweighted.

Network	N	D	PL	CC	AD
PowerGrid	4941	0.0005	19	0.11	2.7
WikiVote	7066	0.004	3.3	0.21	28.5
PGP	10,680	0.0004	7.5	0.44	4.6
Gnutella04	10,876	0.0006	4.6	0.01	7.4
AstroPh	17,903	0.0012	4.2	0.67	22.0
CondMat	21,363	0.0004	5.4	0.70	8.5
HEPTh	27,400	0.0009	4.3	0.34	25.7
Enron	33,696	0.0003	4.0	0.71	10.7
Gnutella31	62,561	0.00008	5.9	0.01	4.7
Epinions	75,877	0.0001	4.3	0.26	10.7
Slashdot	82,168	0.0001	4.1	0.10	12.2
Amazon	262,111	0.00003	8.8	0.43	6.9

Table 1: Network Properties. **Key:** *N= # of nodes, D= density, PL = characteristic path length, CC = local clustering coefficient, AD = average degree.*

4. NETWORK SAMPLING

In the present work, we focus on a particular class of sampling strategies, which we refer to as *link-trace sampling*. In *link-trace sampling*, the next node selected for inclusion into the sample is always chosen from among the set of nodes directly connected to those already sampled. In this way, sampling proceeds by tracing or following links in the network. This concept can be defined formally.

Definition 5. Given an integer k and an initial node (or seed) $v \in V$ to which S is initialized (i.e. $S = \{v\}$), a *link-trace sampling* algorithm, $\mathcal{A}$, is a process by which nodes are iteratively selected from among the current neighborhood $N(S)$ and added to S until $|S| = k$.

Link-trace sampling may also be referred to as *crawling* (since links are "crawled" to access nodes) or viewed as *online* sampling (since the network G reveals itself iteratively during the course of the sampling process). The key advantage of sampling through link-tracing, then, is that complete access to the network in its entirety is *not* required. This is beneficial for scenarios where the network is either large (e.g. an online social network), decentralized (e.g. an unstructured P2P network), or both (e.g. the Web).

As an aside, notice from Definition 5 that we have implicitly assumed that the neighbors of a given node can be obtained by visiting that node during the sampling process (i.e. $N(S)$ is known). This, of course, accurately characterizes most real scenarios. For instance, neighbors of a Web page can be gleaned from the hyperlinks on a visited page and neighbors of an individual in an online social network can be acquired by viewing (or "scraping") the friends list.

Having provided a general definition of *link-trace sampling*, we must now address *which* nodes in $N(S)$ should be preferentially selected at each iteration of the sampling process. This choice will obviously directly affect the properties of the sample being constructed. We study seven different approaches - all of which are quite simple yet, at the same time, ill-understood in the context of real-world networks.

Breadth-First Search (BFS). Starting with a single seed node, the BFS explores the neighbors of visited nodes. At each iteration, it traverses an unvisited neighbor of the *earliest* visited node [14]. In both [30] and [42], it was empirically shown that BFS is biased towards high-degree and high-PageRank nodes. BFS is used prevalently to crawl and collect networks (e.g. [41]).

Depth-First Search (DFS). DFS is similar to BFS, except that, at each iteration, it visits an unvisited neighbor of the most *recently* visited node [14].

Random Walk (RW). A random walk simply selects the next hop uniformly at random from among the neighbors of the current node [37].

Forest Fire Sampling (FFS). FFS, proposed in [34], is essentially a probabilistic version of BFS. At each iteration of a BFS-like process, a neighbor v is only explored according to some "burning" probability p. At $p = 1$, FFS is identical to BFS. We use $p = 0.7$, as recommended in [34].

Degree Sampling (DS). The DS strategy involves greedily selecting the node $v \in N(S)$ with the highest degree (i.e. number of neighbors). A variation of DS was analytically and empirically studied as a P2P search algorithm in [2]. Notice that, in order to select the node $v \in N(S)$ with the highest degree, the process must know $|N(\{v\})|$ for each $v \in N(S)$. That is, knowledge of $N(N(S))$ is required at each iteration. As noted in [2], this requirement is acceptable for some domains such as P2P networks and certain social networks. The DS method is also feasible in scenarios where 1) one is interested in efficiently "downsampling" a network to a connected subgraph, 2) a crawl is repeated and history of the last crawl is available, or 3) the proportion of the network accessed to construct a sample is less important.

SEC (Sample Edge Count). Given the currently constructed sample S, how can we select a node $v \in N(S)$ with the highest degree *without* having knowledge of $N(N(S))$? The SEC strategy tracks the links from the currently constructed sample S to each node $v \in N(S)$ and selects the node v with the most links from S. In other words, we use the degree of v in the induced subgraph of $S \cup \{v\}$ as an approximation of the degree of v in the original network G. Similar approaches have been employed as part of Web crawling strategies with some success (e.g. [9]).

XS (Expansion Sampling). The XS strategy is based on the concept of expansion from work on expander graphs and seeks to greedily construct the sample with the maximal expansion: $\text{argmax}_{S:\,|S|=k} \frac{|N(S)|}{|S|}$, where k is the desired sample size [23, 40]. At each iteration, the next node v selected for inclusion in the sample is chosen based on the expression:

$$\underset{v \in N(S)}{\text{argmax}} |N(\{v\}) - (N(S) \cup S)|.$$

Like the DS strategy, this approach utilizes knowledge of $N(N(S))$. In Sections 5.3 and 6.2, we will investigate in

detail the effect of this expansion bias on various properties of constructed samples.

5. EVALUATING REPRESENTATIVENESS

What makes one sampling strategy "better" than another? In computer science, "better" is typically taken to be structural *representativeness* (e.g. see [25, 29, 35]). That is, samples are considered better if they are more representative of structural properties in the original network. There are, of course, numerous structural properties from which to choose, and, as correctly observed by Ahmed et al. [4], it is not always clear which should be chosen. Rather than choosing arbitrary structural properties as measures of representativeness, we select specific measures of representativeness that we view as being potentially useful for real applications. We divide these measures (described below) into three categories: Degree, Clustering, and Reach. For each sampling strategy, we generate 100 samples using randomly selected seeds, compute our measures of representativeness on each sample, and plot the average value as sample size grows. (Standard deviations of computed measures are discussed in Section 5.4. Applications for these measures of representativeness are discussed later in Section 7.) Due to space limitations and the large number of networks evaluated, for each evaluation measure, we only show results for two datasets that are illustrative of general trends observed in all datasets. However, full results are available as supplementary material.[2]

5.1 Degree

The degrees (numbers of neighbors) of nodes in a network is a fundamental and well-studied property. In fact, other graph-theoretic properties such as the average path length between nodes can, in some cases, be viewed as byproducts of degree (e.g. short paths arising from a small number of highly-connected hubs that act as conduits [5]). We study two different aspects of degree (with an eye towards real-world applications, discussed in Section 7).

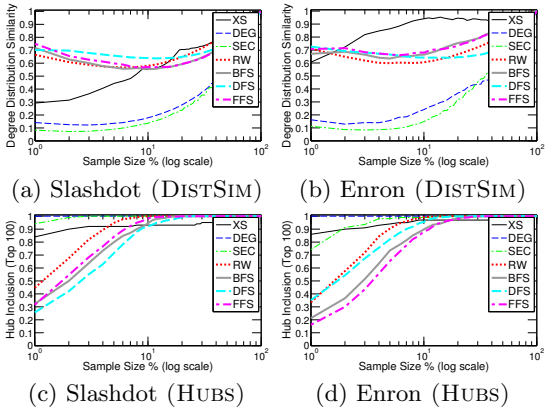
5.1.1 Measures

Degree Distribution Similarity (DistSim). We take the degree sequence of the sample and compare it to that of the original network using the two-sample Kolmogorov-Smirnov (K-S) D-statistic [34], a distance measure. Our objective here is to measure the agreement between the two degree distributions in terms of both shape and location. Specifically, the D-statistic is defined as $D = \max_x\{|F(x) - F_S(x)|\}$, where x is the range of node degrees, and F and F_S are the cumulative degree distributions for G and G_S, respectively [34]. We compute the distribution similarity by subtracting the K-S distance from one.

Hub Inclusion (Hubs). In several applications, one cares less about matching the *overall* degree distribution and more about accumulating the highest degree nodes into the sample quickly (e.g. immunization strategies [13]). For these scenarios, sampling is used as a tool for information retrieval. Here, we evaluate the extent to which sampling strategies accumulate hubs (i.e. high degree nodes) quickly into the sample. As sample size grows, we track the proportion of the top K nodes accumulated by the sample. For our tests, we use $K = 100$.

5.1.2 Results

Figure 1 shows the *degree distribution similarity* (DistSim) and *hub inclusion* (Hubs) for the Slashdot and Enron datasets. Note that the SEC and DS strategies, both of which are biased to high degree nodes, perform best on *hub inclusion* (as expected), but are the *worst* performers on the DistSim measure (which is also a direct result of this bias). (The XS strategy exhibits a similar trend but to a slightly lesser extent.) On the other hand, strategies such as BFS, FFS, and RW tend to perform better on DistSim, but worse on Hubs. For instance, the DS and SEC strategies locate the majority of the top 100 hubs with sample sizes less than 1% in some cases. BFS and FFS require sample sizes of over 10% (and the performance differential is larger when locating hubs ranked higher than 100). More importantly, no strategy performs best on *both* measures. This, then, suggests a tension between goals: constructing small samples of the most well-connected nodes is in conflict with producing small samples exhibiting representative degree distributions. More generally, when selecting sample elements, choices resulting in gains for one area can result in losses for another. Thus, these choices must be made in light of how samples will be used - a subject we discuss in greater depth in Section 7. We conclude this section by briefly noting that the trend observed for SEC seems to be somewhat dependent upon the quality and number of hubs actually present in a network (relative to the size of the network, of course). That is, SEC matches DS more closely as degree distributions exhibit longer and denser tails (as shown in Figure 2). We will revisit this in Section 6.3. (Other strategies are sometimes affected similarly, but the trend is much less consistent.) In general, we find SEC best matches DS performance on many of the social networks (as opposed to technological networks such as the PowerGrid with few "good" hubs, lower average degree, and longer path lengths). However, further investigation is required to draw firm conclusions on this last point.



(a) Slashdot (DistSim) (b) Enron (DistSim)

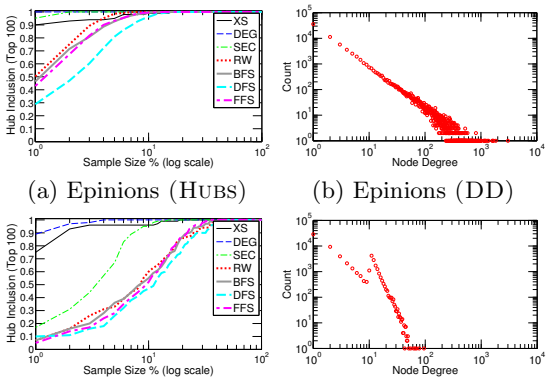
(c) Slashdot (Hubs) (d) Enron (Hubs)

Figure 1: Evaluating (DistSim) and (Hubs). Results for remaining networks are similar.

5.2 Clustering

Many real-world networks, such as social networks, exhibit a much higher level clustering than what one would expect at random [50]. Thus, clustering has been another graph property of interest for some time. Here, we are interested

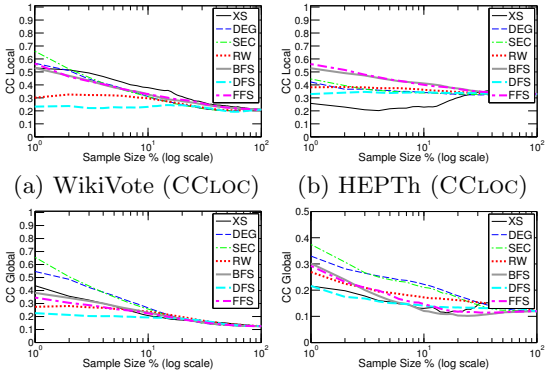
[2]Supplementary material for this paper is available at: `http://arun.maiya.net/papers/supp-netbias.pdf`



(a) Epinions (HUBS) (b) Epinions (DD)

(c) Gnutella31 (HUBS) (d) Gnutella31 (DD)

Figure 2: Performance of SEC on HUBS (shown in green on left) is observed to be dependent on the tail of the degree distributions (DD). SEC matches DS more closely when more and better quality hubs are present. For HUBS, SEC generally performs best on the social networks evaluated.



(a) WikiVote (CCLOC) (b) HEPTh (CCLOC)

(c) WikiVote (CCGLB) (d) HEPTh (CCGLB)

Figure 3: Evaluating CCGLB and CCLOC.

in evaluating the extent to which samples exhibit the level of clustering present in the original network. We employ two notions of clustering, which we now describe.

5.2.1 Measures

Local Clustering Coefficient (CCLOC)**.** The local clustering coefficient [43] of a node captures the extent to which the node's neighbors are also neighbors of each other. Formally, the local clustering coefficient of a node is defined as $C_L(v) = \frac{2\ell}{d_v(d_v-1)}$ where d_v is the degree of node v and ℓ is the number of links among the neighbors of v. The average local clustering coefficient for a network is simply $\frac{\sum_{v \in V} C_L(v)}{|V|}$.

Global Clustering Coefficient (CCGLB)**.** The global clustering coefficient [43] is a function of the number of triangles in a network. It is measured as the number of closed triplets divided by the number of connected triples of nodes.

5.2.2 Results

Results for clustering measures are less consistent than for other measures. Overall, DFS and RW strategies appear to fare relatively better than others. We do observe that, for many strategies and networks, estimates of clustering are initially higher-than-actual and then gradually decline (see Figure 3). This agrees with intuition. Nodes in clusters should intuitively have more paths leading to them and will, thus, be encountered earlier in a sampling process (as opposed to nodes not embedded in clusters and located in the periphery of a network). This, then, should be taken into consideration in applications where accurately matching clustering levels is important.

5.3 Network Reach

We propose a new measure of representativeness called *network reach*. As a newer measure, *network reach* has obviously received considerably less attention than Degree and Clustering within the existing literature, but it is, nevertheless, a vital measure for a number of important applications (as we will see in Section 7). *Network reach* captures the extent to which a sample *covers* a network. Intuitively, for a sample to be truly representative of a large network, it should consist of nodes from diverse portions of the network, as opposed to being relegated to a small "corner" of the graph. This concept will be made more concrete by discussing in detail the two measures of *network reach* we employ: *community reach* and the *discovery quotient*.

5.3.1 Measures

Community Reach (CNM **and** RAK**).** Many real-world networks exhibit what is known as *community structure*. A *community* can be loosely defined as a set of nodes more densely connected among themselves than to other nodes in the network. Although there are many ways to represent community structure depending on various factors such as whether or not overlapping is allowed, in this work, we represent community structure as a *partition*: a collection of disjoint subsets whose union is the vertex set V [18]. Under this representation, each subset in the partition represents a community. The task of a community detection algorithm is to identify a partition such that vertices within the same subset in the partition are more densely connected to each other than to vertices in other subsets [18]. For the criterion of *community reach*, a sample is more representative of the network if it consists of nodes from more of the communities in the network. We measure *community reach* by taking the number of communities represented in the sample and dividing by the total number of communities present in the original network. Since a community is essentially a cluster of nodes, one might wonder why we have included *community reach* as a measure of *network reach*, rather than as a measure of *clustering*. The reason is that we are slightly less interested in the structural details of communities detected here. Rather, our aim is to assess how "spread out" a sample is across the network. Since community detection is somewhat of an inexact science (e.g. see [21]), we measure *community reach* with respect to two separate algorithms. We employ both the method proposed by Clauset et al. in [12] (denoted as CNM) and the approach proposed by Raghavan et al. in [45] (denoted as RAK). Essentially, for our purposes, we are defining communities simply as the output of a community detection algorithm.

Discovery Quotient (DQ)**.** An alternative view of *network reach* is to measure the proportion of the network that is *discovered* by a sampling strategy. The number of nodes discovered by a strategy is defined as $|S \cup N(S)|$. The *discovery quotient* is this value normalized by the total number of nodes in a network: $\frac{|S \cup N(S)|}{|V|}$. Intuitively, we are defin-

ing the *reach* of a sample here by measuring the extent to which it is one hop away from the rest of the network. As we will discuss in Section 7, samples with high *discovery quotients* have several important applications. Note that a simple greedy algorithm for coverage problems such as this has a well-known sharp approximation bound of $1-1/e$ [16,39]. However, link-trace sampling is restricted to selecting subsequent sample elements from the current neighborhood $N(S)$ at each iteration, which results in a much smaller search space. Thus, this approximation guarantee can be shown not to hold within the context of link-trace sampling.

5.3.2 Results

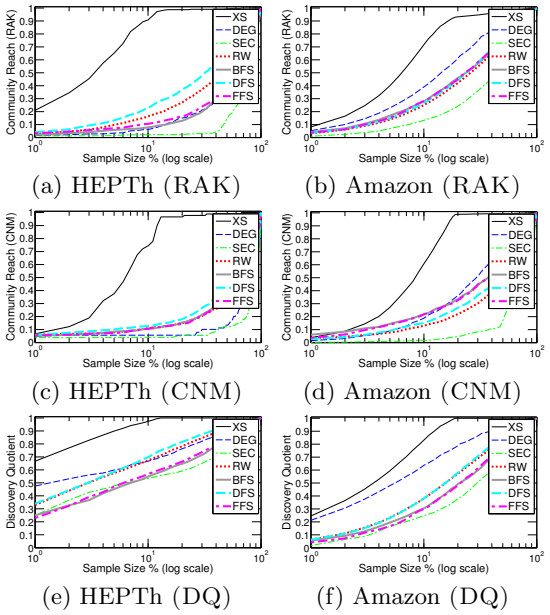
As shown in Figure 4, the XS strategy displays the overwhelmingly best performance on all three measures of *network reach*. We highlight several observations here. First, the extent to which the XS strategy outperforms all others on the RAK and CNM measures is quite striking. We posit that the expansion bias of the XS strategy "pushes" the sampling process towards the inclusion of new communities not already seen (see also [40]). In Section 6.2, we will analytically examine this connection between expansion bias and *community reach*. On the other hand, the SEC method appears to be among the least effective in reaching different communities or clusters. We attribute this to the fact that SEC preferentially selects nodes with many connections to nodes already sampled. Such nodes are likely to be members of clusters already represented in the sample. Second, on the DQ measure, it is surprising that the DS strategy, which explicitly selects high degree nodes, often fails to even come close to the XS strategy. We partly attribute this to an overlap in the neighborhoods of well-connected nodes. By explicitly selecting nodes that contribute to *expansion*, the XS strategy is able to discover a much larger proportion of the network in the same number of steps - in some cases, by actively sampling comparatively *lower* degree nodes. Finally, it is also surprising that the BFS strategy, widely used to crawl and explore online social networks (e.g [41]) and other graphs (e.g. [42]), performs quite dismally on all three measures. In short, we find that nodes contributing most to the expansion of the sample are unique in that they provide specific and significant advantages over and above those provided by nodes that are simply well-connected and those accumulated through standard BFS-based crawls. These and previously mentioned results are in contrast to the conventional wisdom followed in much of the existing literature (e.g. [2, 13, 30, 41, 42]).

5.4 A Note on Seed Sensitivity

As described, link-trace sampling methods are initiated from randomly selected seeds. This begs the question: How sensitive are these results to the seed supplied to a strategy? Figure 5 shows the standard deviation of each sampling strategy for both *hub inclusion* and *network reach* as sample size grows. We generally find that methods with the most explicit biases (XS, SEC, DS) tend to exhibit the least seed sensitivity and variability, while the remaining methods (BFS, DFS, FFS, RW) exhibit the most. This trend is exhibited across all measures and all datasets.

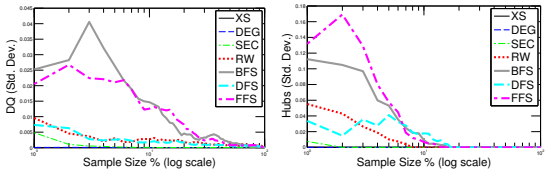
6. ANALYZING SAMPLING BIASES

Let us briefly summarize two main observations from Section 5. We saw that the XS strategy dramatically outperformed all others in accumulating nodes from many different communities. We also saw that the SEC strategy was often a reasonably good approximation to directly sampling high degree nodes and locates the set of most well-connected nodes significantly faster than most other methods. Here, we turn our attention to analytically examining these observed connections. We begin by briefly summarizing some existing analytical results.



(a) HEPTh (RAK) (b) Amazon (RAK)
(c) HEPTh (CNM) (d) Amazon (CNM)
(e) HEPTh (DQ) (f) Amazon (DQ)

Figure 4: Evaluating *network reach*. Results for remaining networks are similar with XS exhibiting superior performance on all three criteria.



(a) Epinions (DQ) (b) Epinions (HUBS)

Figure 5: Standard deviation for DQ and HUBS on Epinions network. Results are similar for remaining networks.

6.1 Existing Analytical Results

Random Walks (RW). There is a fairly large body of research on random walks and Markov chains (see [37] for an excellent survey). A well-known analytical result states that the probability (or *stationary* probability) of residing at any node v during a random walk on a connected, undirected graph converges with time to $\frac{d_v}{2\cdot|E|}$, where d_v is the degree of node v [37]. In fact, the *hitting time* of a random walk (i.e. the expected number of steps required to reach a node beginning from any node) has been analytically shown to be directly related to this stationary probability [24]. Random walks, then, are naturally biased towards high degree (and high PageRank) nodes, which provides some theoretical explanation as to why RW performs slightly better than other strategies (e.g. BFS) on measures such as *hub inclusion*. However, as shown in Figure 1, it is nowhere near

the best performers. Thus, these analytical results appear only to hold in the limit and fail to predict actual sampling performance.

Degree Sampling (DS). In studying the problem of searching peer-to-peer networks, Adamic et al. [2] proposed and analyzed a greedy search strategy very similar to the DS sampling method. This strategy, which we refer to as a degree-based walk, was analytically shown to quickly find the highest-degree nodes and quickly cover large portions of scale-free networks. Thus, these results provide a theoretical explanation for performance of the DS strategy on measures such as *hub inclusion* and the *discovery quotient*.

Other Results. As mentioned in Section 2, to the best of our knowledge, much of the other analytical results on sampling bias focus on *negative* results [1, 15, 30, 31, 46]. Thus, these works, although intriguing, may not provide much help in the way of explaining *positive* results shown in Section 5.

We now analyze two methods for which there are little or no existing analytical results: XS and SEC.

6.2 Analyzing XS Bias

A widely used measure for the "goodness" or the strength of a community in graph clustering and community detection is *conductance* [26], which is a function of the fraction of total edges emanating from a sample (lower values mean stronger communities):

$$\varphi(S) = \frac{\sum_{i \in S, j \in \overline{S}} a_{ij}}{\min(a(S), a(\overline{S}))}$$

where a_{ij} are entries of the adjacency matrix representing the graph and $a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij}$, which is the total number of edges incident to the node set S.

It can be shown that, provided the conductance of communities is sufficiently low, sample expansion is directly affected by community structure. Consider a simple random graph model with vertex set V and a community structure represented by partition $C = \{C_1, \ldots, C_{|C|}\}$ where $C_1 \cup \ldots \cup C_{|C|} = V$. Let e_{in} and e_{out} be the number of each node's edges pointing within and outside the node's community, respectively. These edges are connected uniformly at random to nodes either within or outside a node's community, similar to a configuration model (e.g., [11]). Note that both e_{in} and e_{out} are related directly to conductance. When conductance is lower, e_{out} is smaller[3] as compared to e_{in}. The following theorem expresses the link between expansion and *community reach* in terms of these inward and outward edges.

THEOREM 1. *Let S be the current sample, v be a new node to be added to S, and n be the size of v's community. If $e_{out} < \frac{|V|(e_{in})^2}{n(|V|+e_{in}|S|)}$, then the expected expansion of $S \cup \{v\}$ is higher when v is in a new community than when v is in a current community.*

[3]Suppose conductance of a vertex set Y is $\varphi(Y)$, the total number of edges incident to Y is e, and e_{in} and e_{out} are random variables denoting the inward and outward edges, respectively, of each node (as opposed to constant values). Then, $\mathbb{E}(e_{out}) = \frac{e\varphi(Y)}{|Y|}$ and $\mathbb{E}(e_{in}) = \frac{2e(1-\varphi(Y))}{|Y|}$. If $\varphi(Y) < \frac{2}{3}$, then $\mathbb{E}(e_{out}) < \mathbb{E}(e_{in})$. (In this example, the expectations are over nodes in Y only.)

PROOF. Let X_{new} be the expected value for $|N(\{v\}) - N(S) \cup S|$ when v is in a new community and let X_{curr} be the expected value when not. We compute an upper bound on X_{curr} and a lower bound on X_{new}.

Deriving X_{curr}: Assume v is affiliated with a current community already represented by at least one node in S. Since we are computing an upper bound on X_{curr}, we assume there is exactly one node from S within v's community, as this is the minimum for v's community to be a *current* community. By the linearity of expectations, the upper bound on X_{curr} is $e_{out} + \frac{(n-e_{in})e_{in}}{n}$, where the term $\frac{(n-e_{in})e_{in}}{n}$ is the expected number of nodes in v's community that are both linked to v *and* in the set $V - (N(S) \cup S)$.

Deriving X_{new}: Assume v belongs to a new community not already represented in S. (By definition, no nodes in S will be in v's community.) Applying the linearity of expectations once again, the lower bound on X_{new} is $e_{in} - e_{out}|S|\frac{e_{in}}{|V|}$, where the term $e_{out}|S|\frac{e_{in}}{|V|}$ is the expected number of nodes in v's community that are both linked to v *and* already in $N(S)$.

Solving for e_{out}, if $e_{out} < \frac{|V|(e_{in})^2}{n(|V|+e_{in}|S|)}$, then $X_{new} > X_{curr}$. □

Theorem 1 shows analytically the link between expansion and community structure - a connection that, until now, has only been empirically demonstrated [40]. Thus, a theoretical basis for performance of the XS strategy on *community reach* is revealed.

6.3 Analyzing SEC Bias

Recall that the SEC method uses the degree of a node v in the induced subgraph $G_{S \cup \{v\}}$ as an estimation for the degree of v in G. In Section 5, we saw that this choice performs quite well in practice. Here, we provide theoretical justification for the SEC heuristic. Consider a random network G with some arbitrary expected degree sequence (e.g. a power law random graph under the so-called $G(\mathbf{w})$ model [11]) and a sample $S \subset V$. Let $d(\cdot, \cdot)$ be a function that returns the expected degree of a given node in a given random network (see [11] for more information on *expected* degree sequences). Then, it is fairly straightforward to show the following holds.

PROPOSITION 1. *For any two nodes $v, w \in N(S)$, if $d(v, G) \geq d(w, G)$, then $d(v, G_{S \cup \{v\}}) \geq d(w, G_{S \cup \{w\}})$.*

PROOF. The probability of an edge between any two nodes i and j in G is $\frac{d(i,G) \cdot d(j,G)}{\Delta}$ where $\Delta = \sum_{m \in V} d(m, G)$. Let $\delta = d(v, G_{S \cup \{v\}}) - d(w, G_{S \cup \{w\}})$. Then,

$$\delta = \sum_{x \in S} \frac{d(x,G) \cdot d(v,G)}{\Delta} - \sum_{x \in S} \frac{d(x,G) \cdot d(w,G)}{\Delta} \quad (1)$$

$$= (d(v,G) - d(w,G)) \sum_{x \in S} \frac{d(x,G)}{\Delta} \quad (2)$$

Since $\delta \geq 0$ only when $d(v, G) \geq d(w, G)$, the proposition holds. □

Combining Proposition 1 with analytical results from [2] (described in Section 6.1) provides a theoretical basis for observed performance of the SEC strategy on measures such as *hub inclusion*. Finally, recall from Section 5.1.2 that the

extent to which SEC matched the performance of DS on HUBS seemed to partly depend on the tail of degree distributions. Proposition 1 also yields insights into this phenomenon. Longer and denser tails allow for more "slack" when deviating from these expectations of random variables (as in real-world link patterns that are not purely random).

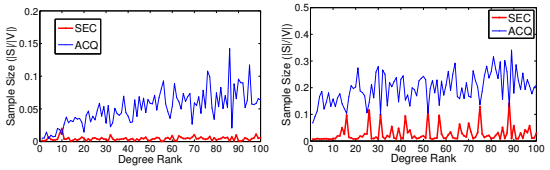
7. APPLICATIONS FOR OUR FINDINGS

We now briefly describe ways in which some of our findings may be exploited in important, real-world applications. Although numerous potential applications exist, we focus here on three areas: 1) Outbreak Detection 2) Landmarks and Graph Exploration 3) Marketing.

7.1 Practical Outbreak Detection

What is the most effective and efficient way to predict and prevent a disease outbreak in a social network? In a recent paper, Christakis and Fowler studied outbreak detection of the H1N1 flu among college students at Harvard University [10]. Previous research has shown that well-connected (i.e. high degree) people in a network catch infectious diseases earlier than those with fewer connections [13, 17, 51]. Thus, *monitoring* these individuals allows forecasting the progression of the disease (a boon to public health officials) and *immunizing* these well-connected individuals (when immunization is possible) can prevent or slow further spread. Unfortunately, identifying well-connected individuals in a population is non-trivial, as access to their friendships and connections is typically not fully available. And, collecting this information is time-consuming, prohibitively expensive, and often impossible for large networks. Matters are made worse when realizing that most existing network-based techniques for immunization selection and outbreak detection assume full knowledge of the global network structure (e.g. [36, 48]). This, then, presents a prime opportunity to exploit the power of *sampling*.

To identify well-connected students and predict the outbreak, Christakis and Fowler [10] employed a sampling technique called *acquaintance sampling* (ACQ) based on the so-called friendship paradox [10, 13, 51]. The idea is that random neighbors of randomly selected nodes in a network will tend to be highly-connected [13, 17, 51]. Christakis and Fowler [10], therefore, sampled random friends of randomly selected students with the objective of constructing a sample of highly-connected individuals. Based on our aforementioned results, we ask: Can we do better than this ACQ strategy? In previous sections, we showed empirically and analytically that the SEC method performs exceedingly well in accumulating hubs. (It also happens to require less information than DS and XS, the other top performers.) Figure 6 shows the sample size required to locate the top-ranked well-connected individuals for both SEC and ACQ. The performance differential is quite remarkable, with the SEC method faring overwhelmingly better in quickly zeroing in on the set of most well-connected nodes. Aside from its superior performance, SEC has one additional advantage over the ACQ method employed by Christakis and Fowler. The ACQ method assumes that nodes in V can be selected uniformly at random. It is, in fact, dependent on this [13]. (ACQ, then, is *not* a link-trace sampling method.) By contrast, SEC, as a pure link-trace sampling strategy, has no such requirement and, thus, can be applied in realistic scenarios for which ACQ is unworkable.



(a) Slashdot (b) CondMat

Figure 6: Comparison of SEC and ACQ on quickly locating well-connected individuals (lower is better). SEC far surpasses ACQ. Results are similar for every network.

7.2 Marketing

Recall from Section 5.3 that a community in a network is a cluster of nodes more densely connected among themselves than to others. Identifying communities is important, as they often correspond to real social groups, functional groups, or similarity (both demographic and not) [18]. The ability to easily construct a sample consisting of members from diverse groups has several important applications in marketing. Marketing surveys often seek to construct stratified samples that collectively represent the diversity of the population [28]. If the attributes of nodes are not known in advance, this can be challenging. The XS strategy, which exhibited the best *community reach*, can potentially be very useful here. Moreover, it has the added power of being able to locate members from diverse groups with absolutely no *a priori* knowledge of demographics attributes, social variables, or the overall community structure present in the network. There is also recent evidence to suggest that being able to construct a sample from many different communities can be an asset in effective word-of-mouth marketing [8]. This, then, represents yet another potential marketing application for the XS strategy.

7.3 Landmarks and Graph Exploration

Landmark-based methods represent a general class of algorithms to compute distance-based metrics in large networks quickly [44]. The basic idea is to select a small sample of nodes (i.e. the landmarks), compute offline the distances from these landmarks to every other node in the network, and use these pre-computed distances at runtime to approximate distances between pairs of nodes. As noted in [44], for this approach to be effective, landmarks should be selected so that they *cover* significant portions of the network. Based on our findings for *network reach* in Section 5.3, the XS strategy overwhelmingly yields the best *discovery quotient* and covers the network significantly better than any other strategy. Thus, it represents a promising landmark selection strategy. Our results for the *discovery quotient* and other measures of *network reach* also yield important insights into how graphs should best be explored, crawled, and searched. As shown in Figure 4, the most prevalently used method for exploring networks, BFS, ranks low on measures of *network reach*. This suggests that the BFS and its pervasive use in social network data acquisition and exploration (e.g. see [41]) should possibly be examined more closely.

8. CONCLUSION

We have conducted a detailed study on sampling biases in real-world networks. In our investigation, we found the BFS, a widely-used method for sampling and crawling networks, to be among the worst performers in both discovering

the network and accumulating critical, well-connected hubs. We also found that sampling biases towards high expansion tend to accumulate nodes that are uniquely different from those that are simply well-connected or traversed during a BFS-based strategy. These high-expansion nodes tend to be in newer and different portions of the network not already encountered by the sampling process. We further demonstrated that sampling nodes with many connections from those already sampled is a reasonably good approximation to sampling high degree nodes. Finally, we demonstrated several ways in which these findings can be exploited in real-world application such as disease outbreak detection and marketing. For future work, we intend to investigate ways in which the top-performing sampling strategies can be enhanced for even wider applicability. One such direction is to investigate the effects of alternating or combining different biases into a single sampling strategy.

9. REFERENCES

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *STOC '05*.

[2] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64(4):046135+, Sept. 2001.

[3] N. K. Ahmed, F. Berchmans, J. Neville, and R. Kompella. Time-based sampling of social network activity graphs. In *MLG '10*.

[4] N. K. Ahmed, J. Neville, and R. Kompella. Reconsidering the Foundations of Network Sampling. In *WIN '10*.

[5] A.-L. Barabasi and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, Oct. 1999.

[6] M. Boguná, R. P. Satorras, A. D. Guilera, and A. Arenas. Models of social networks based on social distance attachment. *Physical Review E*, 70(5):056122+, Nov. 2004.

[7] P. Boldi, M. Santini, and S. Vigna. Do Your Worst to Make the Best: Paradoxical Effects in PageRank Incremental Computations. In *WAW '04*.

[8] T. Cao, X. Wu, S. Wang, and X. Hu. OASNET: an optimal allocation approach to influence maximization in modular social networks. In *SAC '10*.

[9] J. Cho, H. G. Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.

[10] N. A. Christakis and J. H. Fowler. Social Network Sensors for Early Detection of Contagious Outbreaks. *PLoS ONE*, 5(9):e12948+, Sept. 2010.

[11] F. Chung and L. Lu. Connected Components in Random Graphs with Given Expected Degree Sequences. *Annals of Combinatorics*, 6(2):125–145, Nov. 2002.

[12] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111+, Dec. 2004.

[13] R. Cohen, S. Havlin, and D. ben Avraham. Efficient Immunization Strategies for Computer Networks and Populations. *arXiv:cond-mat/0207387v3*, Dec. 2003.

[14] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. McGraw-Hill Science / Engineering / Math, 2nd edition, Dec. 2003.

[15] E. Costenbader. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, Oct. 2003.

[16] U. Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, July 1998.

[17] S. L. Feld. Why Your Friends Have More Friends Than You Do. *The Am. J. of Sociology*, 96(6):1464–1477, 1991.

[18] S. Fortunato. Community detection in graphs. *arXiv:0906.0612v2 [physics.soc-ph]*, Jan. 2010.

[19] O. Frank. *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*, chapter 3. Cambridge University Press, Feb. 2005.

[20] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. A Walk in Facebook: Uniform Sampling of Users in Online Social Networks. *arXiv e-print (arXiv:0906.0060v3)*, Feb. 2011.

[21] B. H. Good, Y.-A. de Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106+, Apr. 2010.

[22] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *WWW '00*.

[23] S. Hoory, N. Linial, and A. Wigderson. Expander Graphs and Their Applications. *Bull. Amer. Math. Soc*, 43, 2006.

[24] J. Hopcroft and D. Sheldon. Manipulation-resistant reputations using hitting time. In *WAW'07*.

[25] C. Hubler, H.-P. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis Algorithms for Representative Subgraph Sampling. In *ICDM '08*.

[26] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, May 2004.

[27] J. Kleinberg. Complex Networks and Decentralized Search Algorithms. In *International Congress of Mathematicians (ICM)*, 2006.

[28] E. D. Kolaczyk. *Statistical Analysis of Network Data*, chapter 5. Springer, 2009.

[29] V. Krishnamurthy, M. Faloutsos, M. Chrobak, J. Cui, L. Lao, and A. Percus. Sampling large Internet topologies for simulation purposes. *Computer Networks*, 51(15):4284–4302, Oct. 2007.

[30] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS. *Arxiv e-print (arXiv:1004.1729v1)*, Apr. 2010.

[31] A. Lakhina, J. W. Byers, M. Crovella, and P. Xie. Sampling Biases in IP Topology Measurements. In *INFOCOM '03*.

[32] S. H. Lee, P. J. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102+, Jan. 2006.

[33] J. Leskovec. Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/.

[34] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06*.

[35] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *KDD '05*.

[36] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD '07*.

[37] L. Lovasz. Random Walks on Graphs: A Survey. *Combinatorics: Paul Erdos is 80*, II, 1994.

[38] L. Lovasz. Very large graphs. *arXiv:0902.0132v1*, Feb. 2009.

[39] A. S. Maiya and T. Y. Berger Wolf. Expansion and search in networks. In *CIKM '10*.

[40] A. S. Maiya and T. Y. Berger-Wolf. Sampling Community Structure. In *WWW '10*.

[41] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07*.

[42] M. Najork. Breadth-first search crawling yields high-quality pages. In *WWW '01*.

[43] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, Mar. 2003.

[44] M. Potamias, F. Bonchi, C. Castillo, and A. Gionis. Fast shortest path distance estimation in large networks. In *CIKM '09*.

[45] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106+, Sept. 2007.

[46] M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *PNAS*, 102(12):4221–4224, Mar. 2005.

[47] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Trans. Netw.*, 17(2):377–390, 2009.

[48] H. Tong, B. A. Prakash, C. Tsourakakis, T. E. Rad, C. Faloutsos, and D. H. Chau. On the Vulnerability of Large Graphs. In *ICDM '10*.

[49] D. Tsoumakos and N. Roussopoulos. Analysis and comparison of P2P search methods. In *InfoScale '06*.

[50] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[51] E. W. Zuckerman and J. T. Jost. What Makes You Think You're so Popular? Self-Evaluation Maintenance and the Subjective Side of the "Friendship Paradox". *Social Psychology Quarterly*, 64(3), 2001.